Students: *Rasul Jasir Dent*
Antonia Claésia da Costa Souza
Course: 902 - Speech Processing
Report

Lexicons and language models for ASR using pocketsphinx
- experiments –

Sections

1. Introduction

For decades, automatic speech recognition systems on three separate but closely related elements: an acoustic model, which maps sounds to phonemes, a pronunciation dictionary, which links orthographic words to groups of phonemes, and a language model, which determines whether a string of words is likely (or admissible). In recent years, end-to-end models rooted in deep-learning have increasingly overtaken this three-part architecture in terms of general performance. However, there is still a use-case for tripartite models. In our experiments, we show that defining a limited lexicon and grammar can translate into very high levels of recognition on closed tasks such as recognizing a sequence of numbers. In particular, we show that using a grammar-based model for the expected number of digits can produce WER as low as 1.2 % (+/- 1.3) on clean speech.

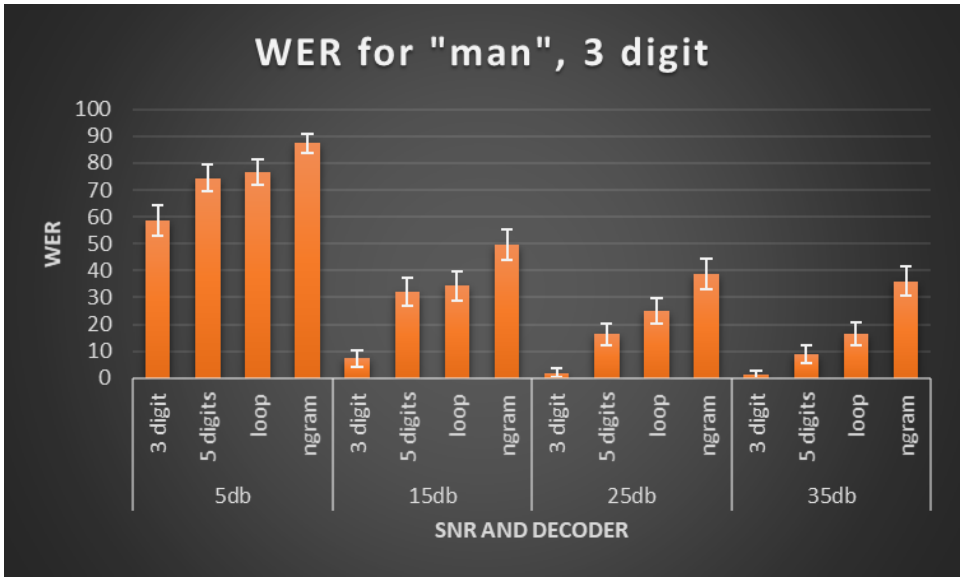
2. Experimental Design

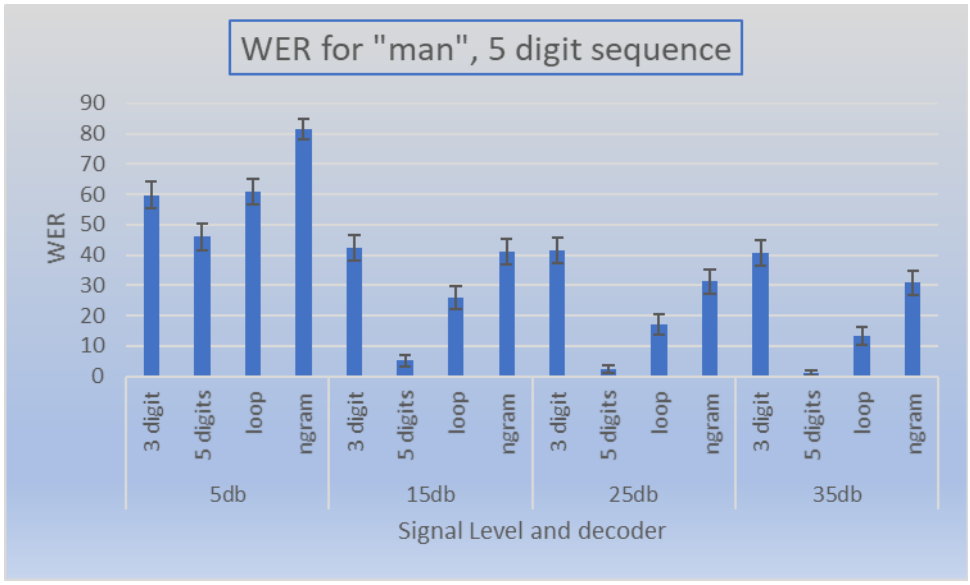
a. Setup: We used the same acoustic model (en-us) for each experiment. For the language model, we tried three different rule-based patterns (5-digit, 3-digit, and loop) and the default ngram model. For the ngram model we used the full CMU pronunciation dictionary, while for the rule based language models, we defined a closed lexicon containing only digits. The relevant entries were taken from the CMU dictionary. The test entries each contained 100 lines of either 3 or 5 words, resulting in a total length of 300 words for the 3-digit tests and 500 words for the 5 digit tests.

b. Experiments: We conducted two principal experiments. First, we examined the impact of language models on recognition, and found that the 3 and 5 digit models had the best accuracy on sequences of their respective lengths. how Signal-to-Noise-Ratio (SNR) and speaker voice affected accuracy. Although all 4 language models were used for experimentation, to test these areas, the most relevant results are those for the 3 and 5 digit models.

3. Results

a. WER by language model and SNR (man) : The first thing to notice is that choosing the correct model is critical for obtaining good results. At a SNR of 5db, all models had high (at least 40%) WER. However, the 3 digit rule and 5 digit rule had the best performance for their respective sequence lengths. At 15db and higher, these two models had WER of 7.3% and 5.2% , which were further reduced to within 2% ± 2% for each at higher SNR.





b. WER by speaker: Since the male voice was the most represented in the test corpus, we predicted that the accuracy would be highest for this speaker. We did not have any defined hypotheses for the other groups. As previously mentioned, the best decoders were the rule-based systems targeted for the sequence length, so for comparison we present the 3-digit decoder on the 3-digit sequence and the 5-digit decoder on the 5-digit sequence.

i. The experiments revealed that the model was most accurate for the male speaker (1.3% +/- 1.3 and 1.2% +/- .95 for 3 and 5 digits, respectively). However, the performance for "woman" was nearly equally good (2.0% +/- 1.59 and 1.2% +/- .95).

ii. In contrast, the results were much worse for the boy (12.67% and 8.8%) girl (13.3% +/- 3.84 and 15.6% +/- 3.18). There are many possible explanations, but this observation is probably tied to children speaking with a higher fundamental frequency, since F0 is one of the most important features.